VAE-Sim: a novel molecular similarity measure based on a variational autoencoder

[1,5]Soumitra Samanta, [2,3,5] Steve O'Hagan, [1]Neil Swainston, [1,6]Timothy J. Roberts & [1,4,*]Douglas B. Kell

[1]Department of Biochemistry and Systems Biology, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Crown St, Liverpool L69 7ZB, UK

[2]Department of Chemistry, [3]The Manchester Institute of Biotechnology, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK.

[4]Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, Building 220, Kemitorvet, 2800 Kgs. Lyngby, Denmark

[5]These authors contributed equally

[6]Present address: University College London Hospital NHS Foundation Trust, 250 Euston Road, London, NW1 2PB.

*Corresponding author dbk@liv.ac.uk

Abstract

Molecular similarity is an elusive but core 'unsupervised' cheminformatics concept, yet different 'fingerprint' encodings of molecular structures return very different similarity values even when using the same similarity metric. Each encoding may be of value when applied to other problems with objective or target functions, implying that *a priori* none is 'better' than the others, nor than encoding-free metrics such as maximum common substructure (MCSS). We here introduce a novel approach to molecular similarity, in the form of a variational autoencoder (VAE). This learns the joint distribution p(z|x) where z is a latent vector and x are the (same) input/output data. It takes the form of a 'bowtie'-shaped artificial neural network. In the middle is a 'bottleneck layer' or latent vector in which inputs are transformed into, and represented as, a vector of numbers (encoding), with a reverse process (decoding) seeking to return the SMILES string that was the input. We train a VAE on over 6 million druglike molecules and natural products (including over one million in the final holdout set). The VAE vector distances provide a rapid and novel metric for molecular similarity that is both easily and rapidly calculated. We describe the method and its application to a typical similarity problem in cheminformatics.

Introduction

The concept of molecular similarity lies at the core of cheminformatics [1-3]. It implies that molecules of 'similar' structure tend to have similar properties. Thus, a typical question can be formulated as follows: "given a molecule of interest M, possibly showing some kind of chemical activity, find me the nearest 50 molecules from a potentially huge online collection to purchase that are most similar to M so I can assess their behaviour in a relevant quantitative-structure-activity (QSAR) analysis".

The most common strategies for assessing molecular similarity involve encoding the molecule as a vector of numbers, such that the vectors encoding two molecules may be compared according to their Euclidean or other distance. In the case of binary strings the Jaccard or Tanimoto similarity (TS) is commonly used [4] as it is a metric (between zero and one). One means for obtaining such a vector for a molecule is to calculate from the structure (or measure) various properties of the molecule ('descriptors' [5-7]), such as clogP or total polar surface area, and then to concatenate them. However, a more common strategy for obtaining the encoding vector of numbers is simply to use structural features directly and to encode them as so-called molecular fingerprints [8-17]. Well-known examples include MACCS [18], atom pairs [19], torsion [20], extended connectivity [21], functional class [22], circular [23], and so on. The similarities so encoded can also then be compared as their Jaccard or Tanimoto similarities. Sometimes a 'difference' or 'distance' is discussed and formulated as 1-TS (a true metric). An excellent and widely used framework for doing all of this is RDKit (www.rdkit.org/) [24], that presently contains nine methods for producing molecular fingerprints.

The problem comes from the fact that the 'most similar' molecules to a target molecule often differ wildly both as judged by their structures observable by eye and quantitatively in terms of the value of the TS of the different fingerprints [25]. As a very small and simple dataset, we take the set of molecules observed by Dickens and colleagues [26] to inhibit the transporter-mediated uptake of the second-generation atypical antipsychotic drug clozapine. These are Olanzapine, Chlorpromazine, Quetiapine, Prazosin, Lamotrigine, Indatraline, Verapamil and Rhein. Of the FDA approved drugs, we assessed the top 50 drugs in terms of their structural similarity to clozapine using the nine RDKit encodings, with the results shown in Table 1 and Fig 1. Only the first four of these are even within the top 50 for any encoding, and only olanzapine appears for each of them. By contrast, the most potent inhibitor is prazosin (which is not even wholly of the same drug class, being both a treatment for anxiety and a high-blood-pressure-lowering agent); however, it appears in the top 50 in only one encoding (torsion) and then with a Tanimoto similarity of just 0.37. This said, visual inspection of their 'Kekularised' structures does show a substantial common substructure between prazosin and clozapine (marked in Fig 1). It is clear that the similarities as judged by standard fingerprint encodings are highly variable, and are prone to both false negatives and false positives when it comes to attacking the question as set down above. What we need is a different kind of strategy.

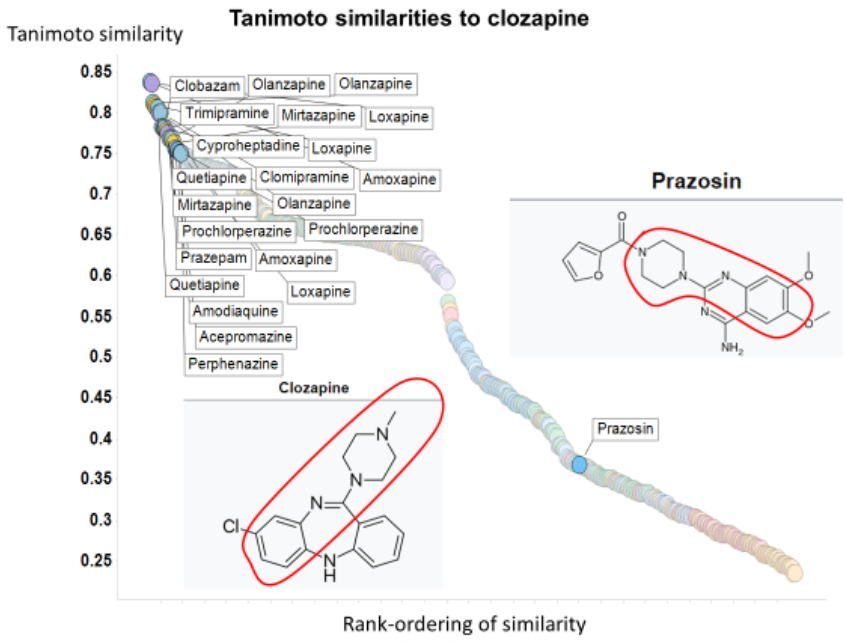
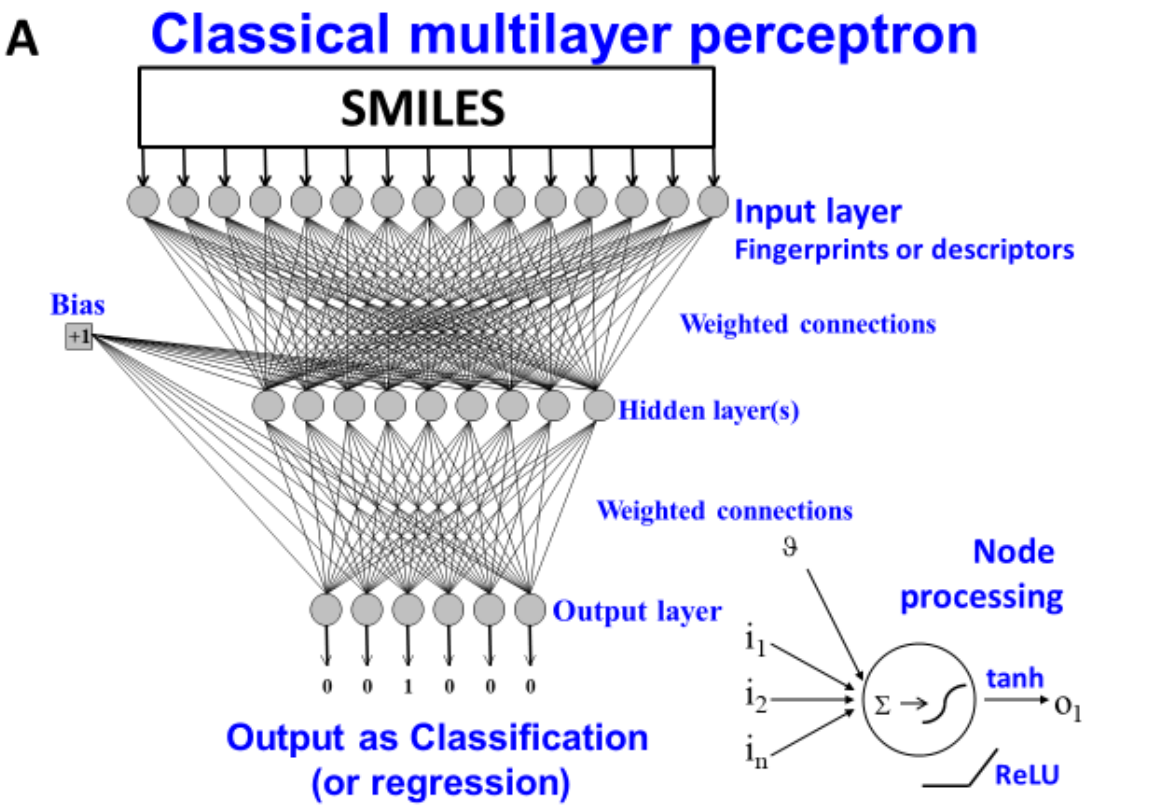
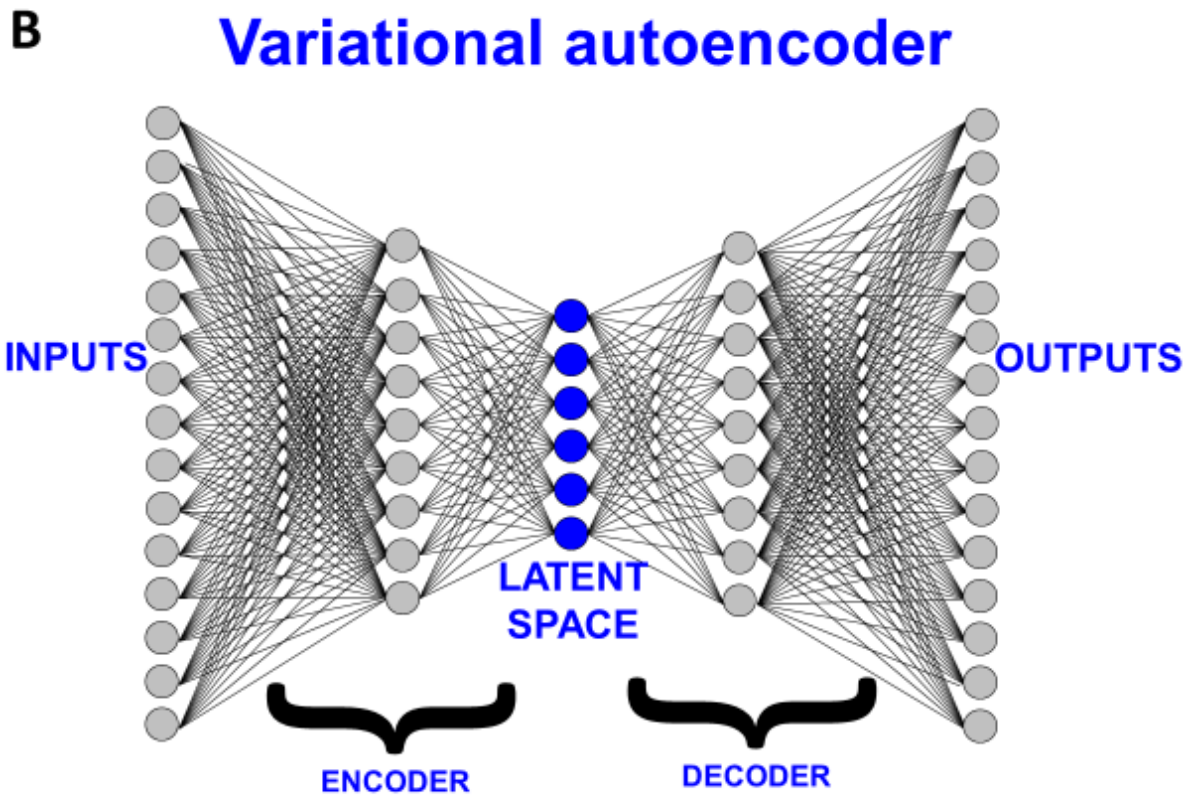


Figure 1. Tanimoto similarities of various molecules to clozapine using the Torsion encoding from RDKit.

Drug	% inhib clozapine uptake	TS Atom Pair	TS Avalon	TS Feat Morgan	TS layered	TS MACCS	TS Morgan	TS Pattern	TS RDKit	TS Torsion
Olanzapine	41	0.68	0.47	0.55	0.77	0.8	0.53	0.81	0.74	0.66
Chlorpromazine	75	0.53		0.35		0.66	0.3	0.74		0.33
Quetiapine	65	0.51	0.57	0.42	0.78		0.35	0.8		0.48
Prazosin	94									0.37
Lamotrigine	26									
Indatraline	35									
Verapamil	83									
Rhein	39									

Table 1. Tanimoto similarity to clozapine using nine different RDKit encodings and their ability to inhibit clozapine transport (data extracted from [26]). A shaded cell means that the molecule was not judged to be in the 'top 50' using that encoding.





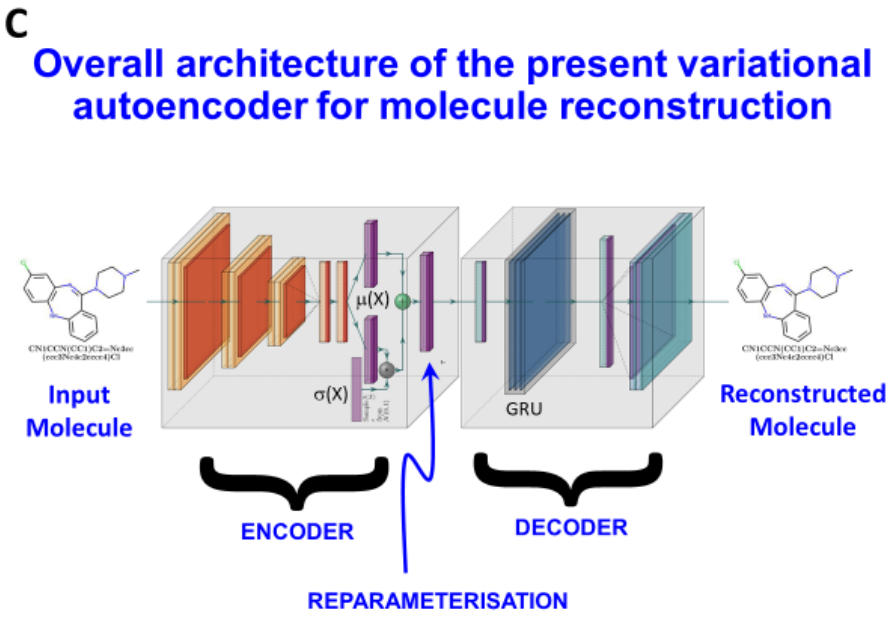


Figure 2. Two kinds of neural architecture. **A**. A classical multilayer perceptron representing a supervised learning system in which molecules encoded as SMILES strings can be used as paired inputs with outputs of interest (whether a classification or a regression). The trained model may then be interrogated with further molecules and the output ascertained. **B**. A variational autoencoder, is a supervised means of fitting distributions of discrete models in a way that reconstructs them via a vector in a latent space. **C**. The VAE architecture used in the present work.

The typical structure of a QSAR type of problem is given in Fig 2A, where a series of molecules represented as SMILES strings [27] are encoded as molecular fingerprints and used to learn a nonlinear mapping to produce an output in the form of a classification or regression estimation. The architecture of this is implicitly in the form of a multilayer perceptron (a classical neural network [28-31]), in which weights are modified ('trained') to provide a mapping from input SMILES to a numerical output. Our fundamental problem stems from the fact that these types of encoding are one-way: The SMILES string can generate the molecular fingerprint but the molecular fingerprint cannot generate the SMILES. Put another way, it is the transition from a world of discrete objects (here molecules) with categorical representations (here SMILES strings) to one of a continuous representation (vectors of weights) that is seemingly irreversible in this representation. One key element is the means by which we can go from a non-numerical representation (such as SMILES or similar [32]) to a numerical representation or 'embedding' (that, as we shall see, is typically constituted by vectors of numbers in the nodes and weights of multilayer neural networks) [33-38]). Deep learning has also been used for the encoding step of 2D chemical structures [39; 40].

More recently, it was recognised that various kinds of architectures could in fact permit the reversal of this numerical encoding so as to return a molecule (or its SMILES string encoding a unique structure). These are known as generative methods [41-50], and at heart their aim to generate a suitable and computationally useful representation [51] of the input data. It is common (but cf. [52; 53]) to contrast two main flavours: generative adversarial networks [54-61] and (especially variational) autoencoders (VAEs) [41; 42; 62-71]. We focus here on the latter, illustrated in Fig 2B.

VAEs are latent-variable generative models that define a joint density $p_\theta(x,z)$ between some observed data $x \in \mathbb{R}^{d_x}$ and unobserved or latent variables $z \in \mathbb{R}^{d_z}$ [72], given some model parameters θ. They use a variational posterior (also referred to as an encoder), $q_\varphi(z \mid x)$, to construct the latent variables with variational parameters φ, and a combination of p(z) and p(x|z) to create a decoder that has the opposite effect. Learning the posterior directly is computationally intractable, so the generic deep learning strategy is to train a neural network to approximate it. The original 'error' backpropagated was based on the Kullback-Leibler (KL) divergence between the desired (log likelihood reconstruction error) and the predicted output distributions [62]. A very great many variants of both architectures and divergence metrics have been proposed since then (not all discernibly better [73]), and it is a very active field (e.g. [58; 59; 74; 75]). Since tuning is necessarily domain-specific [76], and most work is in the processing of images and natural languages rather than in molecules, we merely mention a couple, such as transformers (e.g. [77; 78]) and others (e.g. [79; 80]). Crucial to such autoencoders (that can also be used for data visualisation [81]) is the concept of a bottleneck layer, that as a series of nodes of lower dimensionality than its predecessors or successors, serves to extract or represent [51] the crucial features of the input molecules that are nonetheless sufficient to admit their reconstruction. Indeed, such strategies are sometimes referred to as representational learning.

A higher-level version of the above might state that a good variational autoencoder will project a set of discrete molecules into a continuous latent space represented for any given molecule by the vector representing the values of the outputs of the nodes in the bottleneck layer when it (or its SMILES representation) is applied to the encoder as an input. As with the commonest neural net training system (but cf. [82-86]), we use backpropagation to update the network so as to minimise the difference between the predicted and the desired output, subject to any other constraints that we may apply. We also recognise the importance of various forms of regularisation, that are all designed to prevent overfitting [49; 87-90].

Because the outputs of the nodes in the bottleneck layer both (i) encode the molecule of interest and (ii) effectively represent where molecules are in the chemical space on which they have been trained, a simple metric of similarity between two molecules is clearly the Euclidean or other comparable distance (e.g. cosine distance) between these vectors. This thus provides for a novel type of similarity encoding, that in a sense relates the whole chemical space on which the system has been trained and that we suspect may be of general utility. We might refer to this encoding as the 'essence of molecules' (EM) encoding, but here refer to it as VAE-Sim.

Thus, the purpose of the present article is to describe our own implementation of a simple VAE and its use in molecular similarity measurements as applied, in particular, to the set of drugs, metabolites and natural products that we have been using previously [25; 91-96] as our benchmark for similarity metrics.

Methods

We considered and tested grammar-based and junction-tree methods such as those used by Kajino [35], that exploited some of the ideas developed by Dai [97], Kusner [98] and by Jin and their colleagues [34]. However, our preferred method as described here used one-hot encoding as set out by Gómez-Bombarelli and colleagues [69]. We varied the number of molecules in the training process from ca 250,000 to over 6 million; the large number of possible hyperparameters would have led to a combinatorial explosion, so exhaustive search was (and is) not possible. The final architecture used here (shown in Fig 2C) required 6 days' training on a 1-GPU machine. It involved a CNN encoder with the following layers (Fig 2C): **convolution (1D):** size (in-248=SMILES string length, 40 possible unique SMILES characters, out-9, kernel_size=9), **ReLU,**

convolution (1D): size (in-9, out-9, kernel_size=9) **ReLU, convolution (1D):** size (in-9, out-10, kernel_size=11) **ReLU, Linear (fully connected)**: size(140, latent_dims=100) **SeLU, with VAE mean**- Linear (fully connected): size(140, latent_dims=100) and **variance**- Linear (fully connected): size(140, latent_dims=100). For the decoder we used a **Reparametarization** (combined mean and sigma together) such that the output will be the same as the latent dimension (100 in our case), **Linear (fully connected):** size(latent_dims=100, latent_dims=100) **SeLU, RNN-GRU** (gated neural unit): size (hidden size=488, num_layers=3), **Linear (fully connected):** size(in-488=hidden_gru_size, out-248=SMILES length) **Softmax.** For the loss we used binary cross-entropy + KL-divergence. Neither dropout nor pooling were used. The optimiser was ADAM [99], the fixed learning rate 0.0001, parameters were initialised using the 'Xavier uniform' scheme [100], and a batch size of 128. This was implemented in Python using the Pytorch library (https://pytorch.org/). Most of the pre- and post-processing cheminformatics workflows were written in the KNIME environment (see [101]).

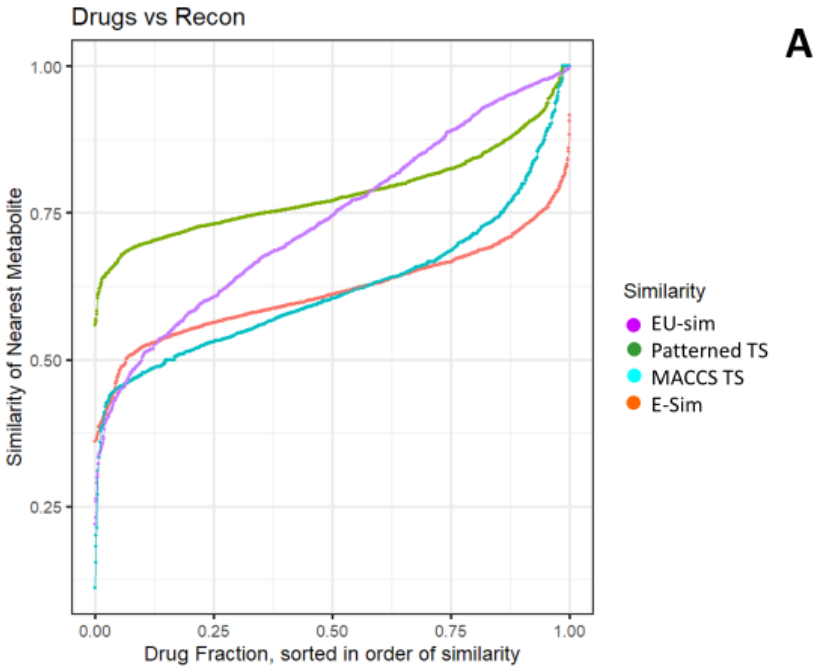
Results

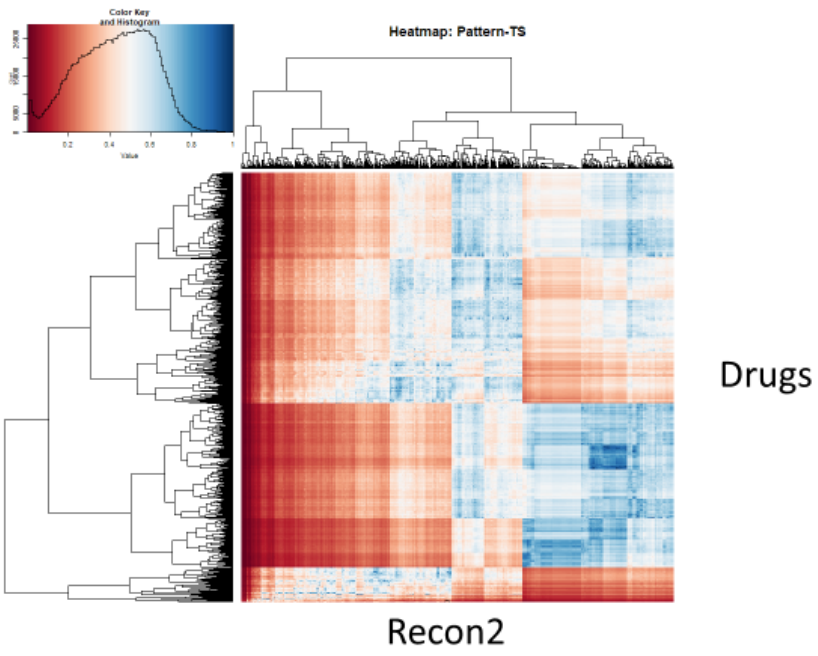
Autoencoders that use SMILES as inputs can return three kinds of outputs: (i) the correct SMILES output mirroring the input and/or translating into the input molecular structure (referred to as 'perfect'), (ii) an incorrect output of a molecule different from the input but that is still legal SMILES (hence will return a valid molecule), referred to as 'good', and (iii) a molecule that is simply not legal SMILES. In practice, our VAE after training returned more than 95% valid SMILES in the test (holdout) set, so those that were invalid could simply be filtered out without significant loss of performance. Following training, each molecule (SMILES) could be associated with a normalised vector of 100 dimensions, and the Euclidean distance between them could be calculated.

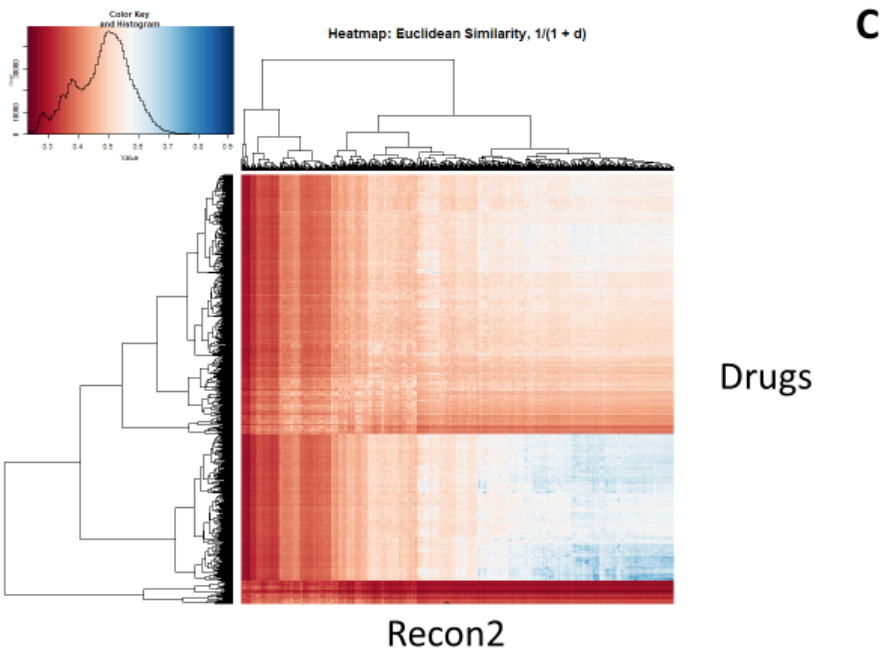
$$E-Sim(x,y) = 1/(1+\sqrt{\sum_{i=1}^{100}(x_i-y_i)^2});\ \ x,y\ \in \mathbb{R}^{100} \quad \text{..... Eq 1}$$

$$EU-Sim(x,y) = 1/(1+\sqrt{\sum_{i=1}^{2}(x_i-y_i)^2});\ \ x,y\ \in \mathbb{R}^{2} \quad \text{..... Eq 2}$$

As previously [25], we compared the similarities between all drugs and all metabolites using the datasets made available in [91]. We here focus on just the MACCS and Patterned encodings of RDKit, and compare them with the normalised Euclidean distances according to the latent vector obtained from the VAE. As before, we rank ordered each drug in terms of its closest similarity to any metabolite. First, Fig 3A (reading from right to left) shows the Tanimoto similarities for the Patterned and MACCS fingerprints, as well as the VAE-Sim values as judged by two metrics. The first, labelled E-Sim (Eq 1), is the Euclidean similarity, based on the raw 100-dimensional hidden vectors, while the second, EU-Sim (Eq 2), used the Uniform Manifold Approximation and Projection (UMAP) dimension reduction algorithm [102; 103] based on the first two UMAP dimensions was used for purposes of visualisation; clearly, as with other encodings, they do not at all follow the same course, and one that may be modified according to the similarity measure used. Figures 3B and 3C show the 'all-vs-all' heatmaps for two of the encodings, indicating again that the VAE-Sim encoding falls away considerably more quickly, i.e. that similarities are judged in a certain sense more 'locally'.







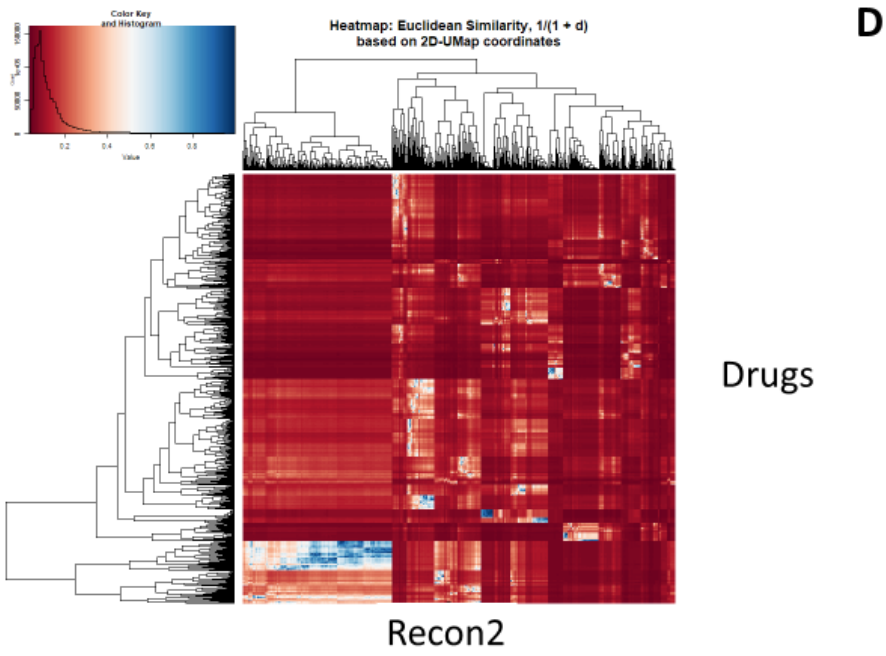


Figure 3. Top similarities between drugs and metabolites as judged by a fingerprint encoding (RDKit patterned) and our new VAE-Sim metric. **A**. Rank ordering. **B**. Heatmap for Tanimoto similarities using RDKit patterned encoding. **C**. Heatmap of Euclidean similarities E-Sim (Eq 1) for VAE-Sim in the 100-dimensional latent vector). **D** Heatmap of Euclidean similarities EU-Sim (Eq 2) for VAE-Sim in 2-dimensional UMAP space.

Figure 4A shows the Patterned similarity for the 'most similar' metabolite for each drug (using TS) compared to that for VAE-Sim (using Euclidean distance), while Figure 4B shows the same for the

MACCS encoding. These again illustrate how the new encoding provides a quite different readout from the standard fingerprint encodings.

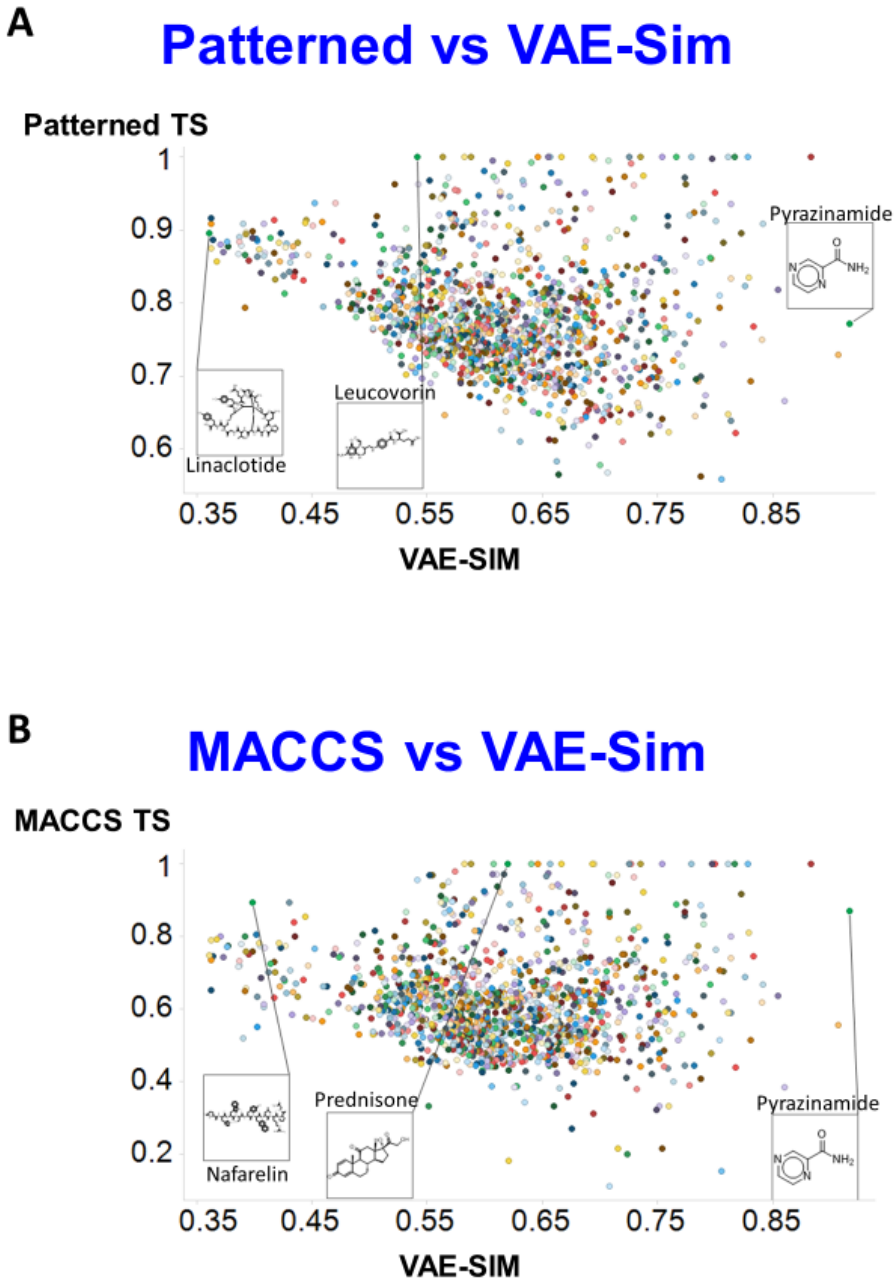


Figure 4. Comparison of similarities between two RDKit fingerprint methods and VAE-Sim Using Tanimoto similarity for fingerprints and Euclidean d_{100} similarity for VAE-Sim. **A**. Patterned encoding. **B**. MACCS encoding.

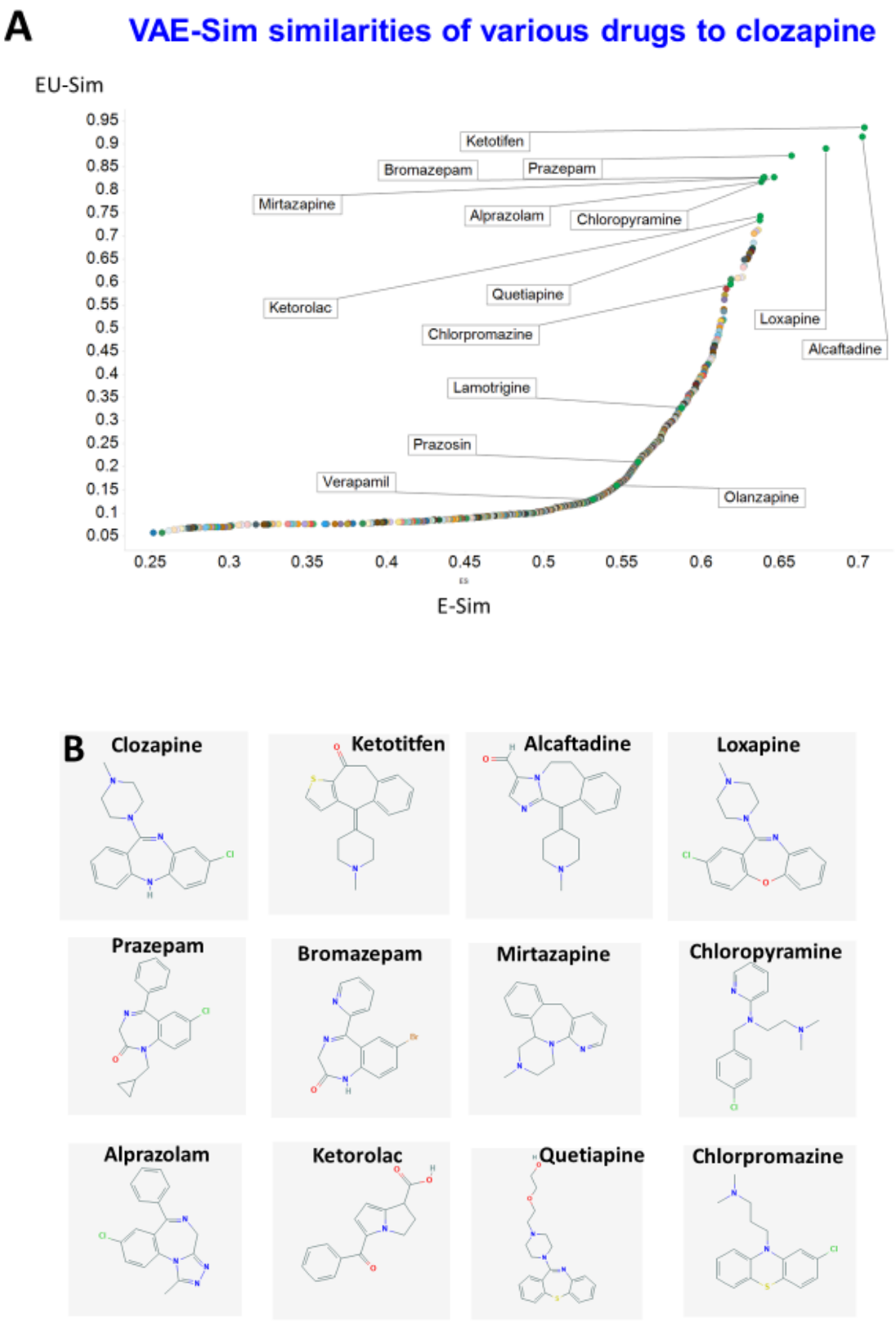


Figure 5. Similarity of drugs to clozapine as judged by the VAE. **A**. Rank order of Euclidean similarity in 100 dimensions (E-Sim) vs 2 UMAP dimensions (EU-Sim) as in Figure 3. Some of the 'most similar' drugs are labelled, as are some of those in Table 1. **B**. Structures of some of the drugs mentioned.

Finally, we used our new metrics to determine the similarity to clozapine of other drugs. Figure 5 shows the two similarity scores based on VAE-Sim, calculated as in Figure 3. Gratifyingly, and while structural similarity is, in part, in the eye of the beholder, a variety of structurally and functionally related antipsychotic drugs such as loxapine, mirtazapine and quetiapine were indeed among the most similar to clozapine, while others not previous considered (such as the antihistamines ketotifen and alcaftadine and the anti-inflammatory COX inhibitor ketorolac) were also suggested as being similar, providing support for the orthogonal utility of the new VAE-Sim metric. However, the rather promiscuous nature of clozapine binding (e.g. [104; 105]), and that of many of the other drugs (e.g. [106-112]), mean that this is not the place to delve deeper.

Discussion

Molecular similarity is at the core of much of cheminformatics (e.g. [3; 8; 113-116]), but is an elusive concept. Our chief interest typically lies in supervised methods such as QSARs, where we use knowledge of paired structures and activities to form a model that allows us to select new structures with potentially desirable activities. Modern modelling methods such as feedforward artificial neural networks based on multilayer perceptrons are very powerful (and they can in fact fit any nonlinear function – the principle of "universal approximation" [117; 118]). Under these circumstances it is usually possible to learn a QSAR using any of the standard fingerprints. However, what we are focused on here is a purely unsupervised representation of the structures themselves (cf [37] which used substructures), and the question of which of these are the 'most similar' to a query molecule of interest. Such unsupervised methods may be taken to include any kinds of unsupervised clustering too (e.g. [119-123]). As with any kind of system of this type, the 'closeness' is a function of the weighting of any individual features, and it is perhaps not surprising that the different fingerprint methods give vastly different similarities, even when judged by rank order (e.g. [25] and above). One similarity measure that is independent of any fingerprint encoding is represented by the maximum common substructure (MCSS). However, by definition, the MCSS uses only part of a molecule; it is also computationally demanding [93; 94], such that 'all-against-all' comparisons such as those presented here are out of the question for large numbers of molecules.

Here we have leveraged a new method that uses only the canonical SMILES encoding of the molecules themselves, leading to its representation as a 100-element vector. Simple Euclidean distances can be used to obtain a metric of similarity that unlike MCSS is rapidly calculated for any new molecule, even against the entire set of molecules used in the development of the latent space.

In addition, unlike any of the other methods described, methods such as VAEs are generative: moving around in the latent space and applying the vector so created to the decoder allows for the generation of entirely new molecules (e.g. [41-45; 48; 50; 58; 60; 68; 69; 124]). This opens up a considerable area of chemical exploration, even in the absence of any knowledge of bioactivities.

What determines the extent to which VAEs can generate novel examples?

The ability of variational autoencoders to generalise is considered to be based on learning a certain 'neighbourhood' around each of the training examples [72; 125], seen as a manifold of lower dimensionality than the dimensionality of the input space [51]. Put another way, "the reconstruction obtained from an optimal decoder of a VAE is a convex combination of examples in the training data" [126]. On this basis, an effect of training set size on the improvement of generalisation (here defined simply as being able to return an accurate answer from a molecule not in the training set) is to be expected, and our ability to generalise (as judged by test set error) improved as the number of molecules increased up to a few million. However, although we did not

explore this, it is possible that our default architecture was simply too large for the smaller number of molecules, as excessive 'capacity' can cause a loss of generalisation ability [126]. This of course leaves open the details of the size and 'closeness' of that neighbourhood, how it varies with the encoding used (our original problem) and what features are used in practice to determine that neighbourhood. The network described here took nearly a week to train on a well-equipped GPU-based machine, and exhaustive analysis of hyperparameters was not possible. Consequently, because an understanding of the importance of local density will vary as a function of the position and nature of the relevant chemical space, we are not going to pursue them here. What is important is (i) that we could indeed learn to navigate these chemical spaces, and (ii) that the VAE approach admits a straightforward and novel estimation of molecular similarity.

Acknowledgments

Present funding includes part of the EPSRC project SuSCoRD (EP/S004963/1), partly sponsored by AkzoNobel. DBK is also funded by the Novo Nordisk Foundation (grant NNF10CC1016517). Tim Roberts is funded by the NIHR UCLH Biomedical Research Centre.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Legends to Figures

Figure 1. Tanimoto similarities of various molecules to clozapine using the Torsion encoding from RDKit.

Figure 2. Two kinds of neural architecture. **A**. A classical multilayer perceptron representing a supervised learning system in which molecules encoded as SMILES strings can be used as paired inputs with outputs of interest (whether a classification or a regression). The trained model may then be interrogated with further molecules and the output ascertained. **B**. A variational autoencoder, is a supervised means of fitting distributions of discrete models in a way that reconstructs them via a vector in a latent space. **C**. The VAE architecture used in the present work.

Figure 3. Top similarities between drugs and metabolites as judged by a fingerprint encoding (RDKit patterned) and our new VAE-Sim metric. **A**. Rank ordering. **B**. Heatmap for Tanimoto similarities using RDKit patterned encoding. **C**. Heatmap of Euclidean similarities E-Sim (Eq 1) for VAE-Sim in the 100-dimensional latent vector). **D** Heatmap of Euclidean similarities EU-Sim (Eq 2) for VAE-Sim in 2-dimensional UMAP space.

Figure 4. Comparison of similarities between two RDKit fingerprint methods and VAE-Sim Using Tanimoto similarity for fingerprints and Euclidean d_{100} similarity for VAE-Sim. **A**. Patterned encoding. **B**. MACCS encoding

Figure 5. Similarity of drugs to clozapine as judged by the VAE. **A**. Rank order of Euclidean similarity in 100 dimensions (E-Sim) vs 2 UMAP dimensions (EU-Sim) as in Figure 3. Some of the 'most similar' drugs are labelled, as are some of those in Table 1. **B**. Structures of some of the drugs mentioned.

References

[1] Gasteiger, J. (2003). Handbook of Chemoinformatics: From Data to Knowledge. Wiley/VCH, Weinheim.

[2] Leach, A. R. & Gillet, V. J. (2007). *An introduction to chemoinformatics, revised edition*. Springer, Dordrecht.

[3] Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. (2014). Molecular Similarity in Medicinal Chemistry. *J Med Chem* **57,** 3186-3204.
[4] Willett, P. (2011). Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. *Wires Data Min Knowl* **1,** 241-251.
[5] Todeschini, R. & Consonni, V. (2009). *Molecular descriptors for cheminformatics, Vol 1. Alphabetical listing*. Wiley-VCH, Weinheim.
[6] Ballabio, D., Manganaro, A., Consonni, V., Mauri, A. & Todeschini, R. (2009). Introduction to MOLE DB - on-line Molecular Descriptors Database. *Match* **62,** 199-207.
[7] Dehmer, M., Varmuza, K. & Bonchev, D. (2012). Statistical modelling of molecular descriptors in QSAR/QSPR. Wiley-VCH, Weinheim.
[8] Bender, A. & Glen, R. C. (2004). Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem* **2,** 3204-18.
[9] Nisius, B. & Bajorath, J. (2010). Rendering conventional molecular fingerprints for virtual screening independent of molecular complexity and size effects. *ChemMedChem* **5,** 859-68.
[10] Owen, J. R., Nabney, I. T., Medina-Franco, J. L. & López-Vallejo, F. (2011). Visualization of molecular fingerprints. *J Chem Inf Model* **51,** 1552-63.
[11] Riniker, S. & Landrum, G. A. (2013). Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *J Cheminform* **5,** 43.
[12] Vogt, M. & Bajorath, J. (2008). Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chem Biol Drug Des* **71,** 8-14.
[13] Awale, M. & Reymond, J. L. (2017). The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J Cheminform* **9,** 11.
[14] Geppert, H. & Bajorath, J. (2010). Advances in 2D fingerprint similarity searching. *Expert Opin Drug Discov* **5,** 529-542.
[15] Muegge, I. & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov* **11,** 137-48.
[16] O'Boyle, N. M. & Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *J Cheminform* **8,** 36.
[17] Willett, P. (2011). Similarity Searching Using 2D Structural Fingerprints. *Meth Mol Biol* **672,** 133-158.
[18] Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* **42,** 1273-80.
[19] Carhart, R. E., Smith, D. H. & Venkataraghavan, R. (1985). Atom Pairs as Molecular-Features in Structure Activity Studies - Definition and Applications. *J Chem Inf Comp Sci* **25,** 64-73.
[20] Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. (1987). Topological Torsion - a New Molecular Descriptor for SAR Applications - Comparison with Other Descriptors. *J Chem Inf Comp Sci* **27,** 82-85.
[21] Rogers, D. & Hahn, M. (2010). Extended-Connectivity Fingerprints. *J Chem Inf Model* **50,** 742-754.
[22] Hassan, M., Brown, R. D., Varma-O'brien, S. & Rogers, D. (2006). Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* **10,** 283-99.
[23] Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S. & Smith, J. (2006). Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *Idrugs* **9,** 199-204.
[24] Riniker, S. & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform* **5,** 26.
[25] O'Hagan, S. & Kell, D. B. (2017). Consensus rank orderings of molecular fingerprints illustrate the 'most genuine' similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important 'natural' drug transporter substrates. *ADMET & DMPK* **5,** 85-125.
[26] Dickens, D., Rädisch, S., Chiduza, G. N., Giannoudis, A., Cross, M. J., Malik, H., Schaeffeler, E., Sison-Young, R. L., Wilkinson, E. L., Goldring, C. E., Schwab, M., Pirmohamed, M. & Nies, A. T. (2018). Cellular uptake of the atypical antipsychotic clozapine is a carrier-mediated process. *Mol Pharm* **15,** 3557-3572.

[27] Weininger, D. (1988). SMILES, a chemical language and information system .1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31-36.
[28] Rumelhart, D. E., McClelland, J. L. & The PDP Research Group. (1986). Parallel Distributed Processing. Experiments in the Microstructure of Cognition, Vols I & II. M.I.T. Press, Cambridge, MA.
[29] Goodacre, R., Kell, D. B. & Bianchi, G. (1993). Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks. *J. Sci. Food Agric.* **63,** 297-307.
[30] Goodacre, R., Timmins, É. M., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B. & Rooney, P. J. (1998). Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology UK* **144,** 1157-1170.
[31] Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V., Radchenko, E., Zefirov, N. S., Makarenko, A. S., Tanchuk, V. Y. & Prokopenko, V. V. (2005). Virtual computational chemistry laboratory - design and description. *J Comput Aided Mol Des* **19,** 453-463.
[32] O'Boyle, N. & Dalke, A. (2018). DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv***,** 7097960.v1.
[33] Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. (2017). Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *arXiv***,** 1701.01329v1.
[34] Jin, W., Barzilay, R. & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv***,** 1802.04364v2.
[35] Kajino, H. (2018). Molecular Hypergraph Grammar with Its Application to Molecular Optimization. *arXiv***,** 1809.02745v1.
[36] Panteleev, J., Gao, H. & Jia, L. (2018). Recent applications of machine learning in medicinal chemistry. *Bioorg Med Chem Lett* **28,** 2807-2815.
[37] Jaeger, S., Fulle, S. & Turk, S. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J Chem Inf Model* **58,** 27-35.
[38] Shibayama, S., Marcou, G., Horvath, D., Baskin, II, Funatsu, K. & Varnek, A. (2020). Application of the mol2vec Technology to Large-size Data Visualization and Analysis. *Mol Inform* **39,** e1900170.
[39] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Adv NIPS 28* **2,** 2224-2232.
[40] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* **30,** 595-608.
[41] Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P. & Schneider, G. (2018). Generative Recurrent Networks for *de novo* drug design. *Mol Inform* **37,** 1700111.
[42] Schneider, G. (2018). Generative models for artificially-intelligent molecular design. *Mol Inform* **37,** 188031.
[43] Grisoni, F. & Schneider, G. (2019). *De novo* Molecular Design with Generative Long Short-term Memory. *Chimia* **73,** 1006-1011.
[44] Arús-Pous, J., Blaschke, T., Ulander, S., Reymond, J. L., Chen, H. & Engkvist, O. (2019). Exploring the GDB-13 chemical space using deep generative models. *J Cheminform* **11,** 20.
[45] Jørgensen, P. B., Schmidt, M. N. & Winther, O. (2018). Deep Generative Models for Molecular Science. *Mol Inform* **37**.
[46] Li, Y., Hu, J., Wang, Y., Zhou, J., Zhang, L. & Liu, Z. (2020). DeepScaffold: A Comprehensive Tool for Scaffold-Based *De Novo* Drug Discovery Using Deep Learning. *J Chem Inf Model* **60,** 77-91.
[47] Lim, J., Hwang, S. Y., Moon, S., Kim, S. & Kim, W. Y. (2020). Scaffold-based molecular design with a graph generative model. *Chem Sci* **11,** 1153-1164.
[48] Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. (2020). Generative molecular design in low data regimes. *Nat Mach Intell* **2,** 171-180.
[49] van Deursen, R., Ertl, P., Tetko, I. V. & Godin, G. (2020). GEN: highly efficient SMILES explorer using autodidactic generative examination networks. *J Cheminform* **12**.

[50] Walters, W. P. & Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. *Nat Biotechnol* **38,** 143-145.
[51] Bengio, Y., Courville, A. & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Trans Patt Anal Machine Intell* **35,** 1798-1828.
[52] Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J. & Schoelkopf, B. (2017). From optimal transport to generative modeling: the VEGAN cookbook. *arXiv***,** 1705.07642.
[53] Husain, H., Nock, R. & Williamson, R. C. (2019). Adversarial Networks and Autoencoders: The Primal-Dual Relationship and Generalization Bounds. *arXiv***,** 1902.00985.
[54] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozairy, S., Courville, A. & Bengio, Y. (2014). Generative Adversarial Nets. *arXiv***,** 1406.2661v1.
[55] Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., Bozdaganyan, M., Aliper, A., Zhavoronkov, A. & Kadurin, A. (2018). Entangled conditional adversarial autoencoder for *de novo* drug discovery. *Mol Pharm* **15,** 4398-4405.
[56] Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein GAN. *arXiv***,** 1701.07875v3.
[57] Goodfellow, I. (2017). Generative Adversarial Networks. *arXiv***,** 1701.00160v1.
[58] Foster, D. (2019). *Generative Deep Learning*. O'Reilly, Sebastopol, CA.
[59] Langr, J. & Bok, V. (2019). *GANs in action*. Manning, Shelter Island, NY.
[60] Prykhodko, O., Johansson, S. V., Kotsias, P. C., Arús-Pous, J., Bjerrum, E. J., Engkvist, O. & Chen, H. M. (2019). A *de novo* molecular generation method using latent vector based generative adversarial network. *J Cheminform* **11**.
[61] Zhao, J. J., Kim, Y., Zhang, K., Rush, A. M. & LeCun, Y. (2017). Adversarially Regularized Autoencoders for Generating Discrete Structures. *arXiv***,** 1706.04223v1.
[62] Kingma, D. & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv***,** 1312.6114v10.
[63] Rezende, D. J., Mohamed, S. & Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv***,** 1401.4082v3.
[64] Doersch, C. (2016). Tutorial on Variational Autoencoders. *arXiv***,** 1606.05908v2.
[65] Benhenda, M. (2017). ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? *arXiv***,** 1708.08227v3.
[66] Griffiths, R.-R. & Hernández-Lobato, J. M. (2017). Constrained Bayesian Optimization for Automatic Chemical Design. *arXiv***,** 1709.05501v5.
[67] Aumentado-Armstrong, T. (2018). Latent Molecular Optimization for Targeted Therapeutic Design. *arXiv***,** 1809.02032.
[68] Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. M. (2018). Application of generative autoencoder in *de novo* molecular design. *Mol Inform* **37,** 1700123.
[69] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. & Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4,** 268-276.
[70] Tschannen, M., Bachem, O. & Lucic, M. (2018). Recent Advances in Autoencoder-Based Representation Learning. 1812.05069v1
[71] Kingma, D. P. & Welling, M. (2019). An Introduction to Variational Autoencoders. *arXiv***,** 1906.02691v1.
[72] Rezende, D. J. & Viola, F. (2018). Taming VAEs. *arXiv***,** 1810.00597v1.
[73] Hutson, M. (2020). Core progress in AI has stalled in some fields. *Science* **368,** 927.
[74] Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G. & Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv* 1804.03599.
[75] Taghanaki, S. A., Havaei, M., Lamb, A., Sanghi, A., Danielyan, A. & Custis, T. (2020). Jigsaw-VAE: Towards Balancing Features in Variational Autoencoders. *arXiv***,** 2005.05496.
[76] Wolpert, D. H. & Macready, W. G. (1997). No Free Lunch theorems for optimization. *IEEE Trans Evol Comput* **1,** 67-82.
[77] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *arXiv***,** 1706.03762.
[78] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv***,** 1810.04805.
[79] Dai, B. & Wipf, D. (2019). Diagnosing and Enhancing VAE Models. *arXiv***,** 1903.05789v2.
[80] Asperti, A. & Trentin, M. (2020). Balancing reconstruction error and Kullback-Leibler divergence in Variational Autoencoders. *arXiv***,** 2002.07514v1.

[81] Goodacre, R., Pygall, J. & Kell, D. B. (1996). Plant seed classification using pyrolysis mass spectrometry with unsupervised learning: The application of auto-associative and Kohonen artificial neural networks. *Chemometr. Intell. Lab. Syst.* **34,** 69-83.
[82] Yao, X. (1999). Evolving artificial neural networks. *Proc. IEEE* **87,** 1423-1447.
[83] Floreano, D., Dürr, P. & Mattiussi, C. (2008). Neuroevolution: from architectures to learning. *Evol Intell* **1,** 47-62.
[84] Vassiliades, V. & Christodoulou, C. (2013). Toward Nonlinear Local Reinforcement Learning Rules Through Neuroevolution. *Neural Computation* **25,** 3020-3043.
[85] Stanley, K. O., Clune, J., Lehman, J. & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nat Mach Intell* **1,** 24-35.
[86] Iba, H. & Noman, N. (2020). Deep Neural Evolution: Deep Learning with Evolutionary Computation. Springer, Berlin.
[87] Le Cun, Y., Denker, J. S. & Solla, S. A. (1990). Optimal brain damage. *Adv Neural Inf Proc Syst* **2,** 598-605.
[88] Dietterich, T. G. (2000). Ensemble methods in machine learning. *LNCS* **1857,** 1-15.
[89] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv***,** 1207.0580.
[90] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv***,** 1609.04836v2
[91] O'Hagan, S., Swainston, N., Handl, J. & Kell, D. B. (2015). A 'rule of 0.5′ for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* **11,** 323-339.
[92] O'Hagan, S. & Kell, D. B. (2015). Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. *Front Pharmacol* **6,** 105.
[93] O'Hagan, S. & Kell, D. B. (2016). MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. *Front Pharmacol* **7,** 266.
[94] O'Hagan, S. & Kell, D. B. (2017). Analysis of drug-endogenous human metabolite similarities in terms of their maximum common substructures. *J Cheminform* **9,** 18.
[95] O'Hagan, S. & Kell, D. B. (2018). Analysing and navigating natural products space for generating small, diverse, but representative chemical libraries. *Biotechnol J* **13,** 1700503.
[96] O'Hagan, S. & Kell, D. B. (2019). Structural similarities between some common fluorophores used in biology and marketed drugs, endogenous metabolites, and natural products. *bioRxiv***,** 834325.
[97] Dai, H., Tian, Y., Dai, B., Skiena, S. & Song, L. (2018). Syntax-directed variational autoencoder for structured data. *arXiv***,** 1802.08786v21.
[98] Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. (2017). Grammar Variational Autoencoder. *arXiv***,** 1703.01925v1.
[99] Kingma, D. P. & Ba, J. L. (2015). ADAM: a method for stochastic optimization. *arXiv***,** 1412.6980v8.
[100] Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proc AISTATs* **9,** 249-256.
[101] O'Hagan, S. & Kell, D. B. (2015). The KNIME workflow environment and its applications in Genetic Programming and machine learning. *Genetic Progr Evol Mach* **16,** 387-391.
[102] McInnes, L., Healy, J. & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv***,** 1802.03426v2.
[103] McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *J Open Source Software* **DOI 10.21105/joss.00861**.
[104] Citraro, R., Leo, A., Aiello, R., Pugliese, M., Russo, E. & De Sarro, G. (2015). Comparative Analysis of the Treatment of Chronic Antipsychotic Drugs on Epileptic Susceptibility in Genetically Epilepsy-prone Rats. *Neurotherapeutics* **12,** 250-262.
[105] Thorn, C. F., Muller, D. J., Altman, R. B. & Klein, T. E. (2018). PharmGKB summary: clozapine pathway, pharmacokinetics. *Pharmacogenet Genomics* **28,** 214-222.
[106] Hopkins, A. L., Mason, J. S. & Overington, J. P. (2006). Can we rationally design promiscuous drugs? *Curr Opin Struct Biol* **16,** 127-36.

[107] Mestres, J., Gregori-Puigjané, E., Valverde, S. & Solé, R. V. (2009). The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol Biosyst* **5,** 1051-7.
[108] Mestres, J. & Gregori-Puigjané, E. (2009). Conciliating binding efficiency and polypharmacology. *Trends Pharmacol Sci* **30,** 470-4.
[109] Oprea, T. I., Bauman, J. E., Bologa, C. G., Buranda, T., Chigaev, A., Edwards, B. S., Jarvik, J. W., Gresham, H. D., Haynes, M. K., Hjelle, B., Hromas, R., Hudson, L., Mackenzie, D. A., Muller, C. Y., Reed, J. C., Simons, P. C., Smagley, Y., Strouse, J., Surviladze, Z., Thompson, T., Ursu, O., Waller, A., Wandinger-Ness, A., Winter, S. S., Wu, Y., Young, S. M., Larson, R. S., Willman, C. & Sklar, L. A. (2011). Drug Repurposing from an Academic Perspective. *Drug Discov Today Ther Strateg* **8,** 61-69.
[110] Dimova, D., Hu, Y. & Bajorath, J. (2012). Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. *J Med Chem* **55,** 10220-8.
[111] Peters, J. U., Hert, J., Bissantz, C., Hillebrecht, A., Gerebtzoff, G., Bendels, S., Tillier, F., Migeon, J., Fischer, H., Guba, W. & Kansy, M. (2012). Can we discover pharmacological promiscuity early in the drug discovery process? *Drug Discov Today* **17,** 325-35.
[112] Hu, Y., Gupta-Ostermann, D. & Bajorath, J. (2014). Exploring compound promiscuity patterns and multi-target activity spaces. *Comput Struct Biotechnol J* **9,** e201401003.
[113] Bajorath, J. (2017). Molecular Similarity Concepts for Informatics Applications. *Methods Mol Biol* **1526,** 231-245.
[114] Eckert, H. & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* **12,** 225-33.
[115] Medina-Franco, J. L. & Maggiora, G. M. (2014). Molecular similarity analysis. In *Chemoinformatics for drug discovery* (ed. J. Bajorath), pp. 343-399. Wiley, Hoboken.
[116] Zhang, B., Vogt, M., Maggiora, G. M. & Bajorath, J. (2015). Comparison of bioactive chemical space networks generated using substructure- and fingerprint-based measures of molecular similarity. *J Comput Aided Mol Des* **29,** 595-608.
[117] Hornik, K., Stinchcombe, M. & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359-366.
[118] Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks* **4,** 251-257.
[119] Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, London.
[120] Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ.
[121] Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis.* Wiley, New York.
[122] Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21,** 3201-3212.
[123] MacCuish, J. D. & MacCuish, N. E. (2011). *Clustering in bioinformatics and drug discovery*. CRC Press, Boca Raton.
[124] Hong, S. H., Ryu, S., Lim, J. & Kim, W. Y. (2020). Molecular Generative Model Based on an Adversarially Regularized Autoencoder. *J Chem Inf Model* **60,** 29-36.
[125] Bozkurt, A., Esmaeili, B., Brooks, D. H., Dy, J. G. & van de Meent, J.-W. (2019). Evaluating Combinatorial Generalization in Variational Autoencoders. *arXiv***,** 1911.04594v1.
[126] Bozkurt, A., Esmaeili, B., Brooks, D. H., Dy, J. G. & van de Meent, J.-W. (2018). Can VAEs Generate Novel Examples? *arXiv***,** 1812.09624v1.